Re-Imagen: Retrieval-Augmented Text-to-Image Generator

Year: 2022 | Citations: 224 | Authors: Wenhu Chen, Hexiang Hu, Chitwan Saharia, William W. Cohen

---

Abstract

Research on text-to-image generation has witnessed significant progress in generating diverse and photo-realistic images, driven by diffusion and auto-regressive models trained on large-scale image-text data. Though state-of-the-art models can generate high-quality images of common entities, they often have difficulty generating images of uncommon entities, such as `Chortai (dog)' or `Picarones (food)'. To tackle this issue, we present the Retrieval-Augmented Text-to-Image Generator (Re-Imagen), a generative model that uses retrieved information to produce high-fidelity and faithful images, even for rare or unseen entities. Given a text prompt, Re-Imagen accesses an external multi-modal knowledge base to retrieve relevant (image, text) pairs and uses them as references to generate the image. With this retrieval step, Re-Imagen is augmented with the knowledge of high-level semantics and low-level visual details of the mentioned entities, and thus improves its accuracy in generating the entities' visual appearances. We train Re-Imagen on a constructed dataset containing (image, text, retrieval) triples to teach the model to ground on both text prompt and retrieval. Furthermore, we develop a new sampling strategy to interleave the classifier-free guidance for text and retrieval conditions to balance the text and retrieval alignment. Re-Imagen achieves significant gain on FID score over COCO and WikiImage. To further evaluate the capabilities of the model, we introduce EntityDrawBench, a new benchmark that evaluates image generation for diverse entities, from frequent to rare, across multiple object categories including dogs, foods, landmarks, birds, and characters. Human evaluation on EntityDrawBench shows that Re-Imagen can significantly improve the fidelity of generated images, especially on less frequent entities.